Journal Of Computer and Information System (J-CIS)
ISSN 2622-0881 (online)

Ekstraksi dan Visualisasi Web Text Mining Menggunakan Jsoup

Sugiarto Cokrowibowo, Ismail

Program Studi Informatika, Universitas Sulawesi Barat
sugiarto.cokrowibowo@unsulbar.ac.id, *ismailmajid@unsulbar.ac.id*

Abstract

Terdapat milyaran dokumen *web* di *world wide web* yang terus bertumbuh dalam volume, kecepatan dan kompleksitas yang besar dan secara alamiah sebagian besar kontennya tidak terstruktur. Diperlukan adanya teknik atau alat untuk mengekstraksi data teks dari sebuah halaman web yang dapat beradaptasi terhadap konten yang tidak terstruktur maupun semi terstruktur dari halaman web. Pada penelitian ini penulis mengajukan pustaka Java Jsoup untuk mengekstraksi dokumen *web* kemudian memvisualisasikan hasilnya dalam bentuk *word cloud*.

Keywords: *web mining, Jsoup, vizualization*

1. Pendahuluan

Web adalah kumpulan dari milyaran dokumen yang sangat besar, beragam, fleksibel dan dinamis. *World Wide Web* terus tumbuh dalam volume, kecepatan, variasi dan kompleksitas yang besar. Hal ini menyebabkan sulitnya mengidentifikasi sebanyak mungkin informasi yang relevan dari *web* yang secara alami sebagian besar kontennya tidak terstruktur. Lahirnya topik *web mining* bertujuan untuk menemukan dan mengekstraksi informasi relevan yang tersembunyi dari dokumen *web* (Jayalatchumy & Thambidurai, 2013).

Web mining adalah proses menemukan dan mengekstraksi mode dan pengetahuan yang berguna dari dokumen dan aktifitas web yang besar menggunakan teknologi data mining (Kosala & Blockeel, 2000). Tujuan utama dari *web mining* adalah mengekstraksi informasi. *Web mining* adalah bagian integrasi dari teknik *data mining* tradisional dengan informasi yang dikumpulkan berasal dari *world wide web*. *Web mining* diuraikan menjadi lima subtugas berikut (Saini & Pandey, 2015):

i. **Resource Discovery**: bagian ini bertugas untuk mengambil dan mengumpulkan informasi dari dokumen-dokumen web yang tidak familiar.

ii. **Information selection and preprocessing**: bagian ini bertugas untuk memilih dan memproses secara otomatis informasi-informasi dari sumber daya web.

iii. **Generalization**: bagian ini bertugas untuk membentuk dan mengungkap pola umum dari situs-situs *web* yang secara alamiah tidak terstruktur/strukturnya beranekaragam.

iv. **Analysis**: bagian ini bertugas memvalidasi dan menginterpretasi pola penambangan data.

v. **Visualization**: bagian ini bertugas memvisualisasikan dan menampilkan data dan informasi hasil *web mining* agar mudah dipahami.

Penelitian relevan pada topik *web mining* yang telah dilakukan sebelumnya diantaranya:

i. Raymond Kosala dan Hendrik Blockeel, pada tahun 2000 mengadakan survey paper berjudul “Web Mining Research: A Survey”. Penelitian ini mengungkapkan topic-topik/ bidang garapan penelitian serta menetapkan klasifikasi awal dari bidang *web mining* yang merupakan bagian integrasi dari *data mining* (Kosala & Blockeel, 2000).

ii. Shipra Saini dan Hari Mohan Pandey, pada tahun 2015 melakukan penelitian berjudul “Review on Web Content Mining Techniques”. Penelitian ini membahas tentang aplikasi-aplikasi dalam *web mining* melalui pendekatan teknik terstruktur, tidak terstruktur, semi terstruktur dan *multimedia data mining* (Saini & Pandey, 2015).

iii. D. Jayalatchumy dan Dr. P. Thamburai, pada tahun 2013 mengadakan survei paper berjudul “Web Mining Research Issue and Future Directions – A Survey”. Penelitian ini memberikan ikhtisar mengenai pengembangan riset di bidang *web mining* serta beberapa isu-isu riset penting yang berhubungan (Jayalatchumy & Thambidurai, 2013).

iv. Pranali Gatfane, Rani Tanpure, Anjali Masodkar dan Vrushali Patil, pada tahun 2015 mengadakan penelitian berjudul “Extraction of Information from Web Page Using Content Mining Approach”. Penelitian ini mengajukan sebuah sistem untuk menghapus berbagai variasi pola derau dari sebuah halaman *web* (Gatfane, Tanpure, Masodkar, & Patil, 2015).

v. Yeqing Li, pada tahun 2017 mengadakan penelitian berjudul “Research on Technology, Algorithm and Application of Web Mining”. Riset ini melakukan analisis realisasi *web content mining* dan *web Structure Mining* serta prinsip-prinsip dasar algoritma dan area aplikasinya (Li, 2017).

vi. Anish Gupta dan Priya Anand, pada tahun 2015 mengadakan penelitian berjudul “FOCUSED WEB CRAWLERS AND ITS APPROACHES”. Penelitian ini mengajukan arsitektur *web crawler* terfokus untuk mengekspos rahasia dibalik implementasi *web crawling* (Gupta & Anand, 2015).

vii. Yaning Yan dan Jing Li, pada tahun 2018 mengadakan penelitian berjudul “Design and Development of an Intelligent Network Crawler System”. Penelitian ini membahas tentang desain dan pengembangan sebuah sistem *crawler* jaringan cerdas menggunakan JAVA WEB (Yan & Li, 2018).

viii. Zejian Shi, Minyong Shi dan Weiguo Lin, pada tahun 2016 mengadakan penelitian berjudul "The Implementation of Crawling News Page Based On Incremental Web Crawler". Penelitian ini mengimplementasikan sebuah filter Bloom yang telah disederhanakan dan hasilnya menunjukkan bahwa *web crawler* dapat memantau halaman berita dengan baik (Shi, Shi, & Lin, 2016).

ix. Sanjay Kumar Malik dan SAM Rizvi, pada tahun 2011 mengadakan penelitian berjudul "Information Extraction using Web Usage Mining, Web Scrapping and Semantic Annotation". Penelitian ini membahas tentang teknik-teknik ekstaksi informasi pada *web* seperti *web usage mining*, *web scrapping* dan *semantic annotation* untuk meningkatkanefisiensi ekstraksi informasi di *web* (Malik & Rizvi, 2011).

x. Chengcheng Hu YingLi, Yongbin Wang dan Lin Wu, pada tahun 2016 mengadakan penelitian berjudul "Analysis of Hot News Based on Big Data". Penelitian ini mendiskusikan beberapa teknologi *web mining* seperti *Scrapy Framework* untuk melakukan *crawl* berita, *Berkeley DB* untuk memfilter URL, algoritma ekstraksi *web text*, teknik untuk melakukan *word segmentation*: *entity recognition tools of natural language processing* dan *weka* untuk menganalisa hasil *scraping* serta visualisasi menggunakan *word clouds* dan diagram. (Hedley)

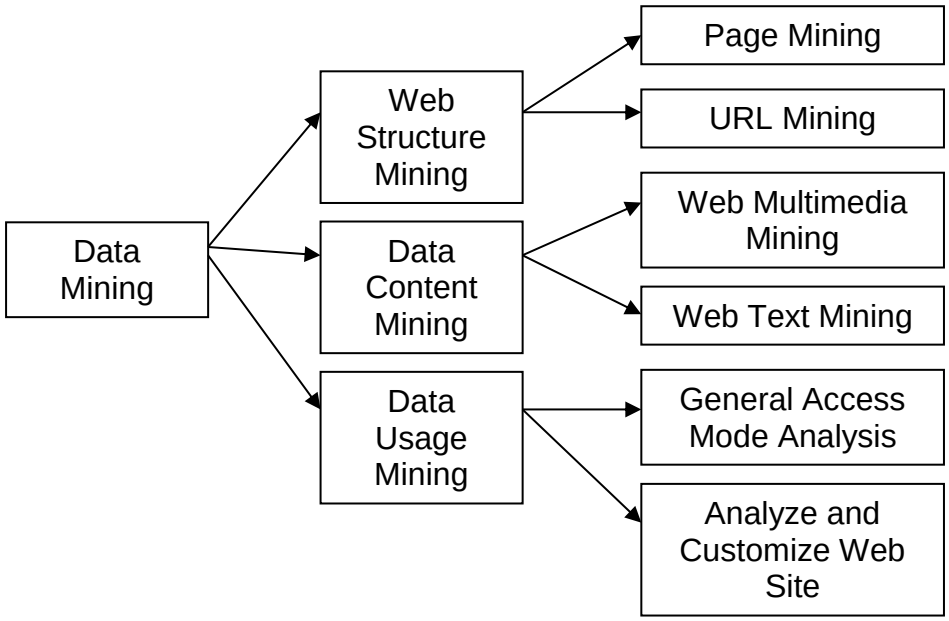
xi. Steffen Lohmann, Florian Heimerl, Fabian Bopp, Michael Burch dan Thomas Ertl, pada tahun 2015 mengadakan penelitian berjudul "ConcentriCloud: Word Cloud Visualization for Multiple Text Documents". Penelitian ini memperkenalkan tentang teknik *Consenri Cloud* untuk membuat visualisasi *word cloud* (Lohmann, Heimerl, Bopp, Burch, & Ertl, 2015).

xii. Zhenfeng He, Ying Cao dan Hui Xiong, pada tahun 2017 mengadakan penelitian berjudul "Generate Galaxy-like Word Cloud Using Molecular Cloud Evolution". Penelitian ini menghasilkan sebuah model dan program *molecular cloud* sederhana untuk membangkitkan *galaxy-like word cloud* (He, Cao, & Xiong, 2017).

Tujuan penelitian ini adalah mendesain dan mengembangkan sistem *web mining* menggunakan Library Jsoup yang termasuk ke dalam klasifikasi *Web Content Mining* khususnya pada bagian *Web Text Mining* selanjutnya hasil ekstraksi menggunakan Jsoup akan divisualisasikan dalam *word cloud* menggunakan JavaFX.

2. Web Mining

Berdasarkan jenis/tipe data yang ditambang dari *web*, *web mining* dibagi menjadi tiga kategori yaitu: *Web Content Mining*, *Web Structure Mining*, dan *Web Usage Mining* (Li, 2017).



Gambar 1. Klasifikasi *Web Mining*

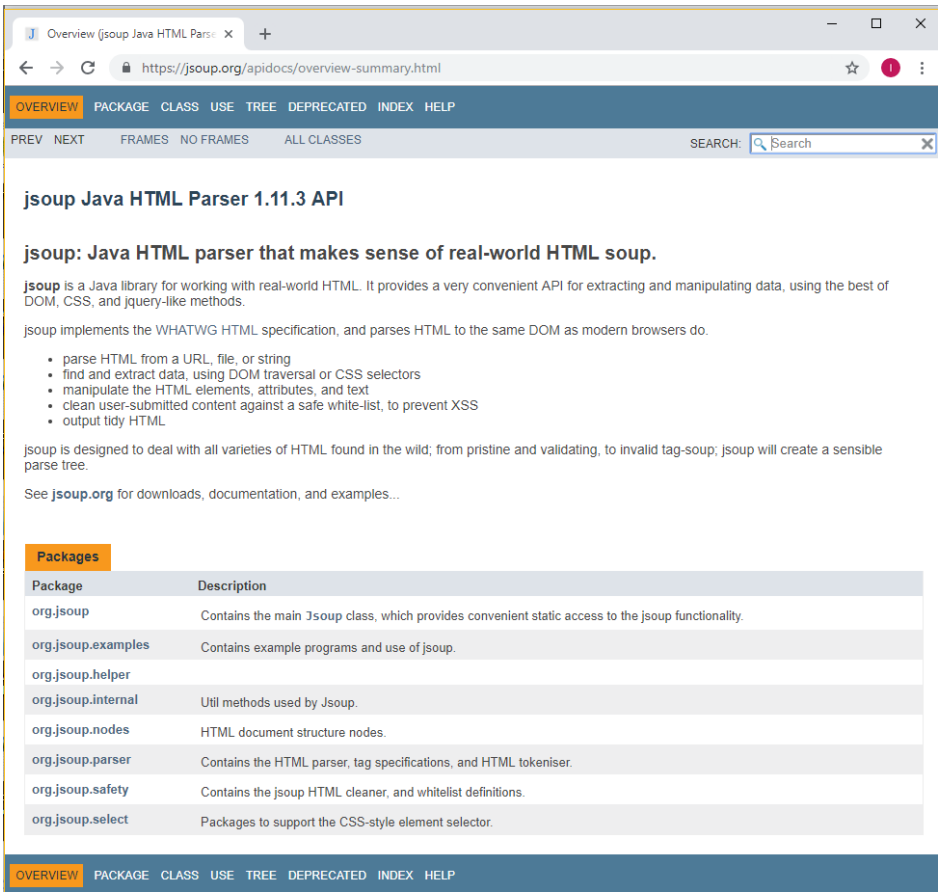
Web Structure Mining terutama berkaitan dengan struktur data web yang dibagi lagi menjadi *Page Structure Mining* dan *URL Mining*. *Web Content Mining* terutama berkaitan dengan data tak terstruktur dan semi terstruktur pada *web* yang dikembangkan melalui topic *Web Text Mining* dan *Web Multimedia Mining*. *Web Usage Mining* dapat dibagi ke dalam dua topic utama *General Access Mode Analysis* dan *Analyze and Customize Web Site* yang bekerja dengan cara menganalisis log-log *website* untuk menemukan beberapa pengetahuan yang berharga (Li, 2017) (Kosala & Blockeel, 2000).

3. Jsoup

Jsoup adalah sebuah pustaka Java untuk bekerja pada dokumen HTML (real-world HTML). Jsoup menyediakan *Application Programming Interface*(API) yang sangat sesuai untuk melakukan ekstraksi dan manipulasi data menggunakan Document Object Model (DOM) terbaik, CSS dan *method-method* yang menyerupai jquery. Jsoup mengimplementasikan spesifikasi WHATWG HTML5 dan memparsing HTML ke DOM yang sama dengan yang ada pada *browser-browser* modern (Hedley, 2018). Berikut ini layanan utama yang tersedia pada pustaka jsoup:

i. Melakukan *scrape* dan *parse* HTML dari sebuah URL, file atau string.
ii. Menemukan dan mengekstraksi data menggunakan DOM *traversal* dan CSS *selector*.
iii. Memanipulasi HTML *elements*, *attributes* dan *text*.
iv. Membersihkan konten yang dikirim oleh pengguna menggunakan *safe white-list* untuk mencegah serangan XSS.
v. Menghasilkan *tidy* HTML.

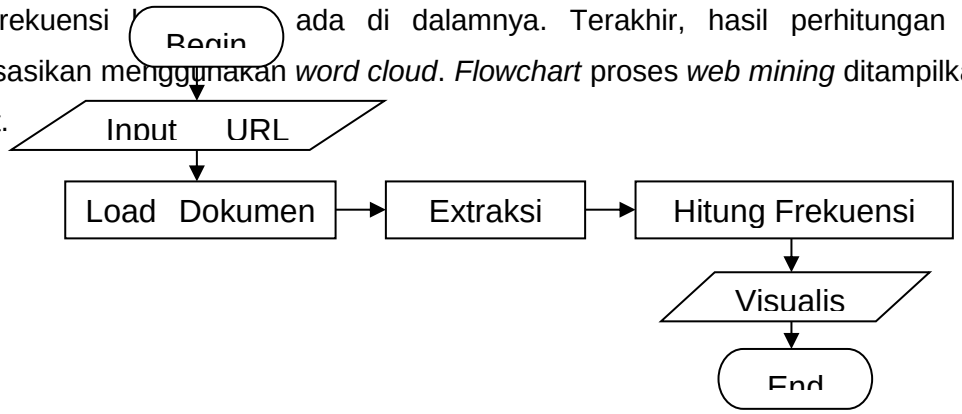
Jsoup didesain untuk dapat menangani semua variasi HTML yang ditemukan, dari memurnikan dan memvalidasi HTML, sampai *tag* jsoup yang tidak valid akan dibuatkan pohon *parse* yang mudah dipahami. Jsoup adalah sebuah proyek *open source* yang dikembangkan oleh Jonathan Hedley dan didistribusikan dibawah lisensi liberal MIT (Hedley, 2018).



Gambar 2: Ikhtisar paket-paket dalam library Jsoup.

4. Desain dan Pengembangan Web Mining

Pada penelitian ini *web mining* akan dikembangkan menggunakan pustaka Jsoup. Sistem akan menerima input berupa URL halaman *web* yang akan diekstraksi. Selanjutnya Jsoup akan meload dokumen berdasarkan input URL. Berikutnya akan dilakukan ekstraksi data dari dokumen web dengan cara memisahkan antara bagian teks dari *tag-tag* HTML. Teks hasil ekstraksi kemudian hitung frekuensi kata yang ada di dalamnya. Terakhir, hasil perhitungan frekuensi akan divisualisasikan menggunakan *word cloud*. *Flowchart* proses *web mining* ditampilkan pada gambar 3 berikut.



Gambar 3: *Flowchart proses*

5. Hasil dan Pembahasan

Implementasi ke dalam source code java untuk *flowchart* ekstraksi dan visualisasi web text mining pada gambar 3 adalah sebagai berikut:

```
private static void miner(String URL) throws IOException {
    //Input URL
    String url = URL;
    //Load Dokumen Web
    Document document = Jsoup.connect(url).get();
    //Ekstraksi data text
    String text = document.text();
    //hitung frekuensi tiap kata
    String[] arrayText = text.trim().split("\\s+");
    ArrayList<Vertex> vertices = hitungFrekuensi(arrayText);
    //nodes = sort(nodes);
    generateWordCloud(vertices);
}
```

Mula-mula sistem akan menerima input URL *web page* yang akan diekstraksi, kemudian dengan bantuan pustaka Jsoup dokumen HTML akan diunduh selanjutnya teks pada *web page* akan diextraksi dan divisualisasikan menggunakan visualisasi *word cloud*. Source code lengkap untuk proses ekstraksi ini dapat dilihat di (Cokrowibowo, 2018). Untuk memvisualisasikan teks hasil ekstraksi halaman *web* akan digunakan teknik visualisasi *word cloud* dengan memberikan variasi pada ukuran kata yang akan ditampilkan berdasarkan frekuensi kemunculan kata tersebut pada halaman *web*.

Hasil pengujian untuk melakukan extraksi terhadap beberapa halaman web diperlihatkan sebagai berikut:

1. URL: http://www.detik.com/

Gambar 4. *Word cloud* hasil ekstraksi URL: http://www.detik.com/

2. URL: https://unsulbar.ac.id/

Gambar 5. *Word cloud* hasil ekstraksi URL: https://unsulbar.ac.id/

3. URL: https://www.themoviedb.org/

Gambar 6. *Word cloud* hasil ekstraksi URL: https://www.themoviedb.org/

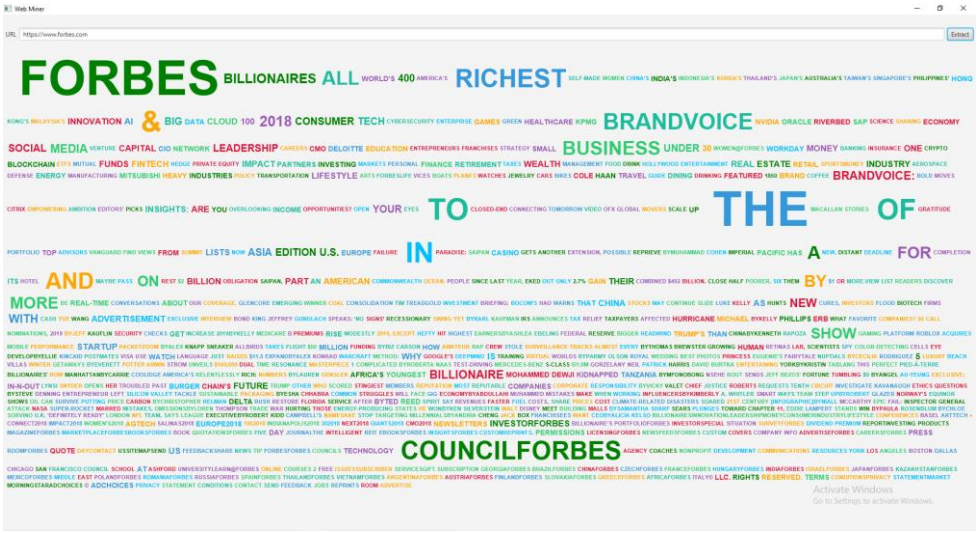
4. URL: https://www.oracle.com/index.html

Gambar 7. *Word cloud* hasil ekstraksi URL: https://www.oracle.com/index.html

5. URL: https://www.tiobe.com/tiobe-index/

Gambar 8. *Word cloud* hasil ekstraksi URL: https://www.tiobe.com/tiobe-index/

6. URL: https://www.forbes.com



Gambar 9. *Word cloud* hasil ekstraksi URL: https://www.forbes.com

6. Penutup

Pada penelitian ini telah dibahas tentang penggunaan pustaka Jsoup untuk mengunduh dan mengekstraksi teks dari halaman web dan kemudian divisualisasikan menggunakan teknik visualisasi *word cloud*. Pengujian untuk ekstraksi dan visualisasi data dari halaman web juga telah dilakukan untuk beberapa sampel URL *website*. Dari lima sub tugas dalam data mining (Resource Discovery, Information selection and preprocessing, Generalization, Analysis, Visualization) penelitian ini hanya menjalankan dua diantaranya yaitu Resource Discovery dan Visualization. Untuk penelitian ke depan kami bermaksud untuk meningkatkan jumlah sub tugas dan tahapan data mining menggunakan teknik-teknik *artificial intelligence* serta mencoba berkolaborasi dengan bidang keilmuan lain yang mempelajari tentang networking dan teori graf serta bidang-bidang ilmu social jika memungkinkan.

REFERENSI

Kosala, R., & Blockeel, H. (2000). Web mining research: a survey. *Acm Sigkdd Explorations Newsletter, 2*(1), 1-5.

Cokrowibowo, S. (2018, 10 15). Retrieved 10 16, 2018, from github: https://gist.github.com/sugiarto-cokrowibowo/94ea942b2a8566b3107584ae46748e83

Gatfane, P., Tanpure, R., Masodkar, A., & Patil, V. (2015). Extaction of Information from Web Page Using Content Mining Approach. *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, 44-48.

Gupta, A., & Anand, P. (2015). FOCUSED WEB CRAWLERS AND ITS APPROACHES. *International Conference on Futuristic Trend in Computational Analysis and Knowledge Management*, 619-622.

He, Z., Cao, Y., & Xiong, H. (2017). Generate Galaxy-like Word Cloud Using Molecular Cloud Evolution. *IEEE. International Conference on Intelligent Human-Machine Systems and Cybernetics*, 77-80.

Hedley, J. (2018, April 15). *jsoup HTML parser*. Retrieved October 13, 2018, from jsoup.org: https://jsoup.org/

Jayalatchumy, D., & Thambidurai, D. (2013). Web Mining Research Issues and Future Directions - A Survey. *IOSR Journal of Computer Engineering, 14*(3), 20-27.

Li, Y. (2017). Research on Technology, Algorithm and Application of Web Mining. *IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, 772-775.

Lohmann, S., Heimerl, F., Bopp, F., Burch, M., & Ertl, T. (2015). ConcentriCloud: Word Cloud Visualization for Multiple Text Documents. *IEEE. International Conference on Information Visualisation*, 114-120.

Malik, S. K., & Rizvi, S. (2011). Information Extraction using Web Usage Mining, Web Scrapping and Semantic Annotation. *IEEE International Conference on Computational Intelligence and Communication System*, 465-468.

Saini, S., & Pandey, H. M. (2015, May). Review on Web Content Mining Techniques. *International Journal of Computer Applications, 118*, 33-36.

Shi, Z., Shi, M., & Lin, W. (2016). The Implementation of Crawling News Page Based On Incremental Web Crawler. *IEEE. Intl Conf on Applied Computing and Information Technology*, 348-351.

Yan, Y., & Li, J. (2018). Design and Development of an Intelligent Network Crawler System. *IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC2018)*, 2667-2670.